Geocoding Addresses from a Large Population-based Study: Lessons Learned

Jane A. McElroy,† Patrick L. Remington,*‡ Amy Trentham-Dietz,*‡ Stephanie A. Robert,§ and Polly A. Newcomb*¶*

Background: Geographic information systems (GIS) and spatial statistics are useful for exploring the relation between geographic location and health. The ultimate usefulness of GIS depends on both completeness and accuracy of geocoding (the process of assigning study participants' residences latitude/longitude coordinates that closely approximate their true locations, also known as address matching). The goal of this project was to develop an iterative geocoding process that would achieve a high match rate in a large population-based health study.

Methods: Data were from a study conducted in Wisconsin using mailing addresses of participants who were interviewed by telephone from 1988 to 1995. We standardized the addresses according to US Postal Service guidelines, used desktop GIS geocoding software and two versions of the Topologically Integrated Geographic Encoding and Referencing street maps, accessed Internet mapping engines for problematic addresses, and recontacted a small number of study participants' households. We also tabulated the project's cost, time commitment, software requirements, and brief notes for each step and their alternatives.

Results: Of the 14,804 participants, 97% were ultimately assigned latitude/longitude coordinates corresponding to their respective residences. The remaining 3% were geocoded to their zip code centroid.

Conclusion: The multiple methods described in this work provide practical information for investigators who are considering the use of GIS in their population health research.

Editor's note: *An invited commentary on this article appears on page 384.*
Submitted 20 August 2002; final version accepted 21 March 2003.

From the *Comprehensive Cancer Center, †Gaylord Nelson Institute for Environmental Studies, ‡Department of Population Health Sciences, and §School of Social Work, University of Wisconsin, Madison, WI; and ¶Cancer Prevention Research Program, Fred Hutchinson Cancer Research Center, Seattle, WA.

This study was supported by National Cancer Institute grants RO1 CA47147 and UO1 CA82004.

Correspondence: Jane A. McElroy, Comprehensive Cancer Center, University of Wisconsin, 610 Walnut St, Room 305 WARF, Madison, WI 53726. E-mail: jamcelroy@wisc.edu.

1044-3983/03/1404-0399
DOI: 10.1097/01.EDE.0000073160.79633.c1

Key Words: geocoding, address matching, data collection, data methods, GIS

(*Epidemiology* 2003;14: 399–407)

Epidemiology has a long history of demonstrating the importance of place in relation to health. With powerful and user-friendly geographic information system (GIS) software, researchers now have more tools to explore the relationship between geographic location and health through mapping.[1,2] In addition, the availability of data collected and compiled by government agencies and the increasing sophistication of inquiry into health determinants allow an unprecedented level of integration of individual, environmental, and community level variables. A critical component for this level of integration in data analysis is complete and accurate geocoding, also known as address matching. The purpose of geocoding is to assign longitude and latitude coordinates to an address by matching the address number to an "address-range" in a digital map (called a street reference map) and, by interpolation, to estimate where the address is located between the two coordinates that define the limits of the address range.[3]

Transforming existing addresses into latitude/longitude coordinates has proved challenging, with match rates as low as 20%.[4] Some addresses may be mapped with low sensitivity because of small deviations in address spelling and specification, changes of street numbering over time, or incomplete street reference maps. Furthermore, addresses such as rural routes or post office boxes cannot be assigned latitude/longitude coordinates using geocoding software because they are not street addresses. Finally, an address match does not necessarily indicate an accurate assignment of latitude/longitude coordinates because of positional inaccuracies in the street reference maps.

We describe here multiple methods used for in-house geocoding that ultimately achieved a 97% match rate to the participants' residences in a large, population-based health study in Wisconsin. In addition, this paper provides practical

information regarding the cost, staffing, time demands, and confidentiality requirements (eg, Health Insurance Portability and Accountability Act) that need to be considered in deciding whether to outsource the geocoding task or complete it in-house.[5–12]

METHODS

Case–Control Study Population

We geocoded participants' mailing addresses from two sequential case–control studies of breast cancer.[13,14] Briefly, all participants (n = 14,804) were English-speaking women living in Wisconsin and between 20 and 79 years of age. We identified incident invasive breast cancer cases from Wisconsin's mandatory statewide tumor registry between 1988 and 1994. Controls were selected randomly from lists obtained from the Wisconsin Department of Transportation (20–64 years of age) and the US Health Care Financing Administration (65–74 years of age). The response rate for participation in the study interview was 85% for cases and 87% for controls.

Geocoding Strategies

The geocoding procedure in this study was a multistep, iterative process with the following two main strategies: (1) to improve the original address quality and (2) to assign latitude/longitude coordinates to the address. Mailing addresses provided by disease surveillance registries, government agencies, or medical facilities are often incomplete or are formatted in a way that is incompatible with the geocoding software used to assign coordinates. This was true for much of the data from this study. We used three steps to improve address quality and two geocoding techniques to assign latitudes/longitude coordinates to participants' residences. For each mailing address, up to five approaches were used (Fig. 1).

Step 1: Address Improvement with Post Office Standardization

The least expensive way to improve mailing addresses was to standardize them to the US Postal Service format. The standardization process changes the mailing addresses into optimal format for delivery by the US Postal Service, which is also optimal for matching with geocoding software. For example, "4345 North 73 Street" is changed to "4345 N 73rd St." (Note: all addresses used as examples have been altered to maintain confidentiality of participants.) If the address was garbled or too abbreviated and the software could not resolve it, these were edited individually (see below). For example, "Wintergreen apt 1001, 5603 Janesville Roa" was changed to

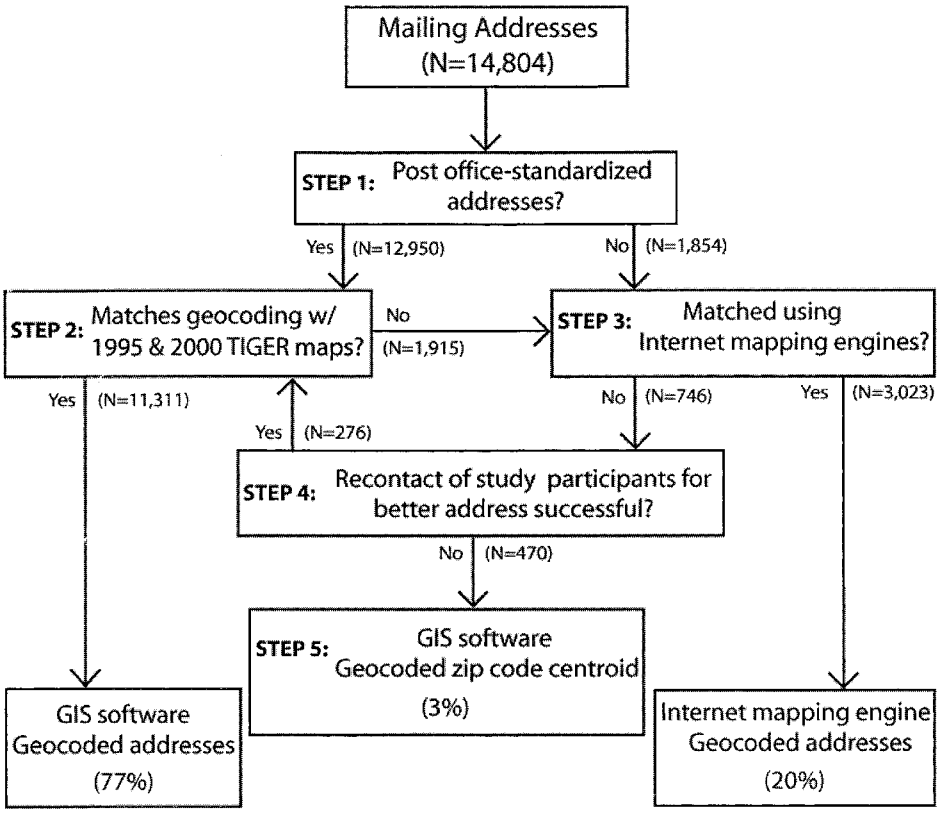

FIGURE 1. The geocoding procedure used to assign latitude/longitude coordinates to 14,804 participants' mailing addresses.

"5603 Janesville Rd." Address standardization can be performed through commercially available software or by submitting the addresses to the Tennessee branch of the US Postal Service, where the Postal Service processes the addresses through their standardization software.

Step 2: Geocoding Using Two Versions of Street Reference Maps

The 2000 (and 1995) Topologically Integrated Geographic Encoding and Referencing (TIGER) street map for Wisconsin was used as the street reference file. This reference map is composed of line segments of the state's street network and the associated numbering system used in addresses.[15] The geocoding software compares the line segments in the street reference map to the participants' addresses to obtain a match between the two. Geocoding software interpolates between line segment nodes (eg, north and south end of a street), estimates the location of the address along the centerline of that street, and subsequently assigns a latitude/longitude coordinate to that point.

Because of the incompleteness of the street reference map and the variable quality of the participants' addresses, we performed a preliminary match of addresses to latitude/longitude coordinates using post office-standardized addresses, two versions of street reference maps, and proprietary geocoding software. Once we completed this initial geocoding procedure, matched addresses were assigned latitude/longitude coordinates and census tract numbers. Geocoding software (ArcView 3.2, Environmental Systems Research Institute, Redlands, CA) assigned census tracts (arranged as polygons) to matched addresses (depicted as points) using the 1990 census tract boundaries. Finally, addresses that did not match were identified. Due to the large number of addresses (nearly 15,000), it was impractical to evaluate each address' quality prior to geocoding.

Participants' street addresses and zip codes were the required elements in the address matching process in ArcView 3.2 software. These addresses were geocoded to an 80% spelling score and 80% overall sensitivity score; these scores allowed matches for addresses with minor deviations in spelling and format. ArcView 3.2 uses a statistical probabilistic approach as its matching strategy. For a disagreement between the street reference map and address, the preprogrammed probability weights for each part of an address penalized the house number the most and the street type the least.[16] Adjusting the spelling and overall sensitivity scores upward, for example, would decrease the chance of placement error but would also increase the number of unmatched records.

Step 3: Address Improvement with Internet Mapping Engines

We used Internet mapping engines for addresses that did not match to a location on either the 1995 or 2000 TIGER street map (Fig. 1). This more resource-intensive address improvement strategy was applied to post office box addresses, rural route addresses, partial or garbled addresses, and unmatched addresses from steps 1 and 2. We entered participants' telephone numbers or names, individually, into an Internet mapping engine (eg, Anywho.com) to identify updated street addresses. Partial, garbled, or unmatched street addresses were edited individually. All updated and edited addresses were entered into an Internet mapping engine to locate the respective latitude/longitude coordinates (eg, Mapblast.com). (Note: MapBlast.com no longer provides latitude/longitude coordinates to mapped locations. TeleAtlas.com, MelissaDATA.com, and the US Census Bureau's Gazetteer currently provide tools to obtain latitude/longitude coordinates from addresses.) For example, "546 HWY B BX 395 E RR,3" was corrected to "546 County Rd B." For the purpose of determining the practical use of intersection information in geocoding, nearest intersections to the address locations were also recorded from the Internet map.

We developed and applied *a priori* criteria for accepting a location that were identified by the Internet engine. For example, if the town, telephone number, and last name were the same as the ones in the address file but a man's first name was listed, an updated street address was recorded. Latitude/longitude coordinates and new location information were entered individually into a customized database.

Step 4: Address Improvement Through Recontacting of Study Participants

If addresses remained unmatched after step 3, we recontacted participants' households by telephone. Improved location information was obtained, including street addresses and the closest street intersection. If a new street address did not match or was not provided using steps 1, 2 or 3, the nearby intersection was used to assign latitude/longitude coordinates (Fig. 1).

Step 5: Geocoding to the Zip Code Centroid

For addresses that failed to match using Steps 2, 3 or 4, the latitude/longitude coordinates of the centroids of the addresses' zip code regions were assigned. The ArcView 3.2 software calculated these latitude/longitude coordinates using the plain geometric mean of East–West and North–South ranges. Other possible options for these nonmatching addresses would include geocoding to the town center, randomly dispersing the nonmatching addresses within the zip code, or eliminating the nonmatching addresses from the analysis.

Hierarchical Rules

When an address had multiple matches, we implemented hierarchical rules to designate a single latitude/longitude coordinate per participant. For recontacted partici-

pants' location information (step 4), matches from street addresses were considered superior to intersections. Matches using the 2000 TIGER street map were considered superior to the 1995 TIGER street map for several reasons. The 2000 TIGER street map had 2% more line segments (694,139 vs. 679,915). Also, the study addresses had been post office-standardized to the current street networks. Due to the temporal nature of the street reference map, a newer street name or numbering expansion might be reflected in 2000 but not in 1995.

Rural and Urban Designation

Data from the US Census for urban and rural categories, standardized to the 1990 population, were used to calculate the percent of each county classified as urban and rural. Although other classification schemes have been developed,[17] we used "percent urban" to represent the percentage of county residents who live in census-defined "urban areas." This includes incorporated cities and villages, as well as census-designated places (eg, census blocks and block groups) with 2,500 or more residents.[18]

RESULTS

Step 1: Address Improvement with Post Office Standardization

Of the original 14,804 addresses, 32% were already in post office-standardized form and 55% were adjusted by the standardization software (Table 1). Tumor registry addresses required the highest percentage of adjustment (58%) followed by addresses from the Wisconsin Department of Transportation (53%) and the US Health Care Financing Administration (52%). Post office standardization identified 1,854 addresses (13%) as unrecognizable. Of these unrecognizable addresses, 507 (27%) were county, state or US highway addresses, 128 (7%) were rural route or post office box addresses, 724 (39%) were garbled or incomplete addresses, and the remaining 495 (27%) had all the necessary components (street range, name and type) but were not recognized. The unrecognizable addresses were more likely to be located in counties with a higher percentage of the population living in rural areas (Pearson correlation 0.41).

Step 2: Geocoding Using Two Versions of Street Reference Maps

By using both the 1995 and 2000 TIGER street maps and post office standardized addresses, 11,311 (76%) addresses were assigned latitude/longitude coordinates (69% with 2000 TIGER and an additional 7% with 1995 TIGER). Over 60% (n = 9,324) of the addresses matched to both the 1995 and 2000 TIGER street maps.

Although the 2000 TIGER street map had more line segments, the 1995 TIGER street map included more unimproved roads, such as those found in northern rural Wisconsin (personal written communication with Catherine Miller, Geography Division, U.S. Census Bureau, 6/24/02). Consequently, the majority of addresses geocoded using the 1995, but not the 2000, TIGER street maps were in rural locations.

Addresses obtained from the US Health Care Financing Administration and the tumor registry had similar match rates (70% and 69% respectively), whereas addresses obtained from Wisconsin Department of Transportation had a lower match rate of 65%. The percentage of addresses in each county that could be geocoded ranged from 0% to 98% (standard deviation = 23%); the median was 48%. This match rate was strongly correlated with the percentage of each county classified as rural (Pearson correlation −0.81; Table 2). The percent of addresses in each county containing post office box and rural route addresses was also strongly correlated with the percent of each county defined as rural (Pearson correlation 0.66).

Step 3: Address Improvement with Internet Mapping Engines

By using Internet mapping engines (eg, MapBlast.com) in conjunction with a database entry system, latitude/longitude coordinates were assigned to 3,023 (80%) of the 3,769 addresses that were not geocoded in Step 2 (Fig. 1).

TABLE 1. Post Office Standardization Results in Relation to the Three Sources of Mailing Addresses

Address source	No.	Addresses Unchanged after Standardization (%)	Addresses Improved after Standardization (%)	Addresses Not Recognized as Valid after Standardization (%)
Cancer Registry*	7248	30	58	12
DOT†	3986	34	53	13
HCFA‡	3570	35	52	13
Total	14804	32	55	13

* Female Wisconsin residents diagnosed with invasive breast cancer.
† Random selection of women 20–64 years of age from driver's license lists provided by the Department of Transportation.
‡ Random selection of women 65–79 years of age from lists provided by the Health Care Financing Administration.

TABLE 2. Successful Geocoding in Relation to Percent of People Living in Rural Areas

Percent Rural*	No. of Counties	Study Participants		Percent Geocoded†	
		No.	(%)	Mean	Range
Counties					
0–30	12	8228	56	99	98–100
31–55	14	2679	18	96	84–100
56–70	14	1843	12	92	71–100
71–99	17	1350	9	87	58–98
100	15	704	5	78	43–97
State					
64	72	14,804	100	90	43–100

* Rural area percent defined according to the US. Census Bureau.[18]
† Assigned latitude/longitude coordinates to study participants' residences.

Step 4: Address Improvement Through Recontact of Study Participants

Recontacting participants by telephone was attempted for 597 unmatched post office box addresses. Surviving participants (70%) or their next of kin (30%) were asked to provide street addresses and nearby intersection street names. Of the 293 (49%) households successfully contacted, 276 provided additional information. Overall, 267 (45%) households provided sufficient information with 90 (34%) addresses and 177 (66%) intersections successfully geocoded. No notable differences in age or educational status of the original study participants were seen between households successfully contacted in comparison with households not successfully contacted or with households that declined to provide adequate information to geocode their addresses. However, significantly more cases (50%) were successfully contacted than controls (37%).

Step 5: Geocoding to Zip Code Centroid

Of the 14,804 original mailing addresses, only 470 addresses (3%) were not assigned latitude/longitude coordinates from steps 2, 3, or 4. Therefore, these subjects were assigned latitude/longitude coordinates corresponding to the centroid of their respective zip code regions. Fourteen percent of these addresses (n = 67) were within minor civil divisions (a city or village).

Geocoding Using Intersection

For 1,059 addresses in which both a new street address and an intersection from reverse telephone or name lookup was recorded (step 3), 79% (n = 840) of the addresses were geocoded within a quarter mile of their geocoded intersection, 90% (n = 951) within half a mile, and 94% (n = 998) within 1 mile.

From additional location information provided by recontacted participants' households, 177 intersections were located on an electronic map, which suggested valid intersection information (step 4). The ability to geocode these intersections varied depending on the street information. Of those with two intersecting streets, 36 of 57 (63%) were geocoded. None of the participants with a street intersecting a county, state, or US highway (n = 71), or two intersecting county, state, or US highways (n = 49) were geocoded.

Costs

The estimated costs for geocoding using the steps described above are listed in Table 3. For our project, we estimate that the total cost was $12,500. Approximately one-half of our expenses were derived from the programmer's time spent on customizing the data entry software (step 3). One quarter of our costs arose from personnel time spent updating street addresses that were problematic (step 3), and another quarter arose from recontacting households (step 4). Steps 1, 2 and 5 had nominal costs.

DISCUSSION

In our large, population-based health study in Wisconsin, we assigned latitude/longitude coordinates to 97% of almost 15,000 participants' residences using the methods detailed above at a cost of less than one dollar per address. Approximately three quarters of the post office-standardized addresses matched to latitude/longitude coordinates using only geocoding software (step 2), which is consistent with other studies.[9,19,20] Higher match rates in urban areas compared with rural areas were also consistent with other reports.[21,22] If we had used only steps 1 and 2 to geocode addresses, participants from rural areas would be more likely to have unknown coordinates. Steps 3 and 4 increased the overall match rate and reduced this urban bias. Other researchers have reported geocoding rates ranging from 20% to 100% depending on factors such as the number of problematic addresses, quality of addresses, and type of street reference map used.[4,9,20,21,23–29]

To minimize costs for in-house geocoding, a high-speed Internet connection and high-quality addresses are essential.[4,30] In addition, personnel familiar with geocoding software and manipulating Internet search engines also improves the efficiency of geocoding. Whether resources should be spent on designing a customized data entry system (step 3) and recontacting participants (step 4) must be evaluated on a project-by-project basis.

Match rates can be improved using multiple approaches. These include additional contact of all participants by telephone or home interviews.[21] Addresses can be marked on local maps[23,31] and on county planning maps[4] by research-

TABLE 3. Resources Needed, Costs Incurred, and Time Required to Geocode

Geocoding Steps	Software and Staff Cost/Time	Access and Software	Notes
Step one: Post office standardization*	≈ 80		
Tennessee Branch of U.S. Postal Service	No charge for first request/2 hour @$20/hour	US Postal Service form 5603: Address File Standardization on Diskette; 800-233-5866	Updates addresses to current year status and flags all changes. No limit on number of addresses; processed in 21 business days
Commercial vendors	$35–$300+/variable	Internet search can provide names of vendors	CASS-certified is the US Postal Service highest standard; confidentiality needs to be assured—might be problematic to outsource addresses given new HIPAA rules and IRB approval requirements
Step two: Geocoding w/TIGER maps*	≈ 380		
In-house geocoding:			
Geocoding software	$300–$1,200	University individual license or desktop software ESRI's ArcView 3.2; www.esri.com	TIGER street map or commercially available street maps can be used with ArcView
Geocoding procedure	1–4 hours@$20/hour		Time varies depending on familiarity of person doing the geocoding with software and reference maps; virtually no time difference is experienced based on number of addresses since the geocoding procedure is done in batch format
Out-sourcing geocoding:			
Commercial vendors	$260–$25,000/variable	Geocoding software package; Internet search can provide names of vendors	See reference #5, 6, 7, 8, 11, 12, 38 for descriptions of software packages and reference maps; time varies according to how vendors deal with post office box, rural route, and garbled addresses. Typically, the least costly price structure geocodes problematic addresses only to zip code centroids or not at all
Reference street maps			
TIGER	Free	www.census.gov/geo/www/cob/bdy_files.html TIGER 1990, 1995 & 2000 maps available	US Census provides these maps in a variety of formats compatible with most geocoding software systems
Commercial vendors	$450–$28,000	Internet search can provide names of vendors	Quality of street maps is variable, especially for rural locations.[34] Typically, these maps are frequently updated
Local (government) agency Project cost	free or nominal fee ≈$380	See Rogers[32] as an example	Available for small geographic areas
Step three: Internet mapping engines*			
On-line directory assistance	Free/≈20 addresses per hour @$20/hour	Examples: www.mapquest.com (uses GDT data); www.mapblast.com (uses GDT data); www.anywho.com (uses mapblast for mapping)	Other mapping engines and name or telephone directories are available; Internet search can provide names of engines; none provide latitude/longitude coordinates
Latitude/longitude providers (examples)			
TeleAtlas	First 100 free; price structure varies; ≈$300 for 15,000	www.teleatlas.com; EZ-Locate software	Requires Microsoft Access version 1997 or older; addresses can be submitted in batch; TeleAtlas provides a statement of non disclosure
US Gazetteer	free	www.census.gov/geo/www/maps/	Click on-line mapping link; uses 1998 TIGER street maps; no street names are provided on map; starts at zip code or city level with zoom-in function
MelissaDATA	free	www.melissadata.com	Click on U.S. Address Lookup; provides zip+4 and latitude/longitude at zip+4 centroid
Customized data entry software and programming	$100–$550; 100 hours @ $55/hours = $5,500	University individual license or desktop software FoxPro 7.0; www.msdn.microsoft.com/vfoxpro	Other data entry software are available, including Oracle and Microsoft Access; Internet search can provide additional names; programmer custom designs data entry system to accommodate multiple geocoding methods. Alternatively, can use a spreadsheet, eg, Microsoft Excel to track updated addresses but this does not permit custom error checks and relational databases

(*Table continues*)

TABLE 3. *Continued*

Geocoding Steps	Software and Staff Cost/Time	Access and Software	Notes
Step four: Recontact participants*	≈ 3,000		
Subjects recontacted by telephone	Two households/ hour @ $20/ hour	Research staff to telephone households	Costs varies depending on telephone or mail contact
US Postal Service postmaster contacted for rural route addresses	Five postmasters/ hour @ $20/ hour	Research staff to telephone postmaster	Requires preliminary work to identify unmatched rural routes and the corresponding post office; might need to develop a data agreement depending on confidentiality issues
Step five: Geocode to zip code centroid*			
Geocoding software	$300–$1,200/	University individual license or desktop software ESRI's ArcView 3.2; www.esri.com	See step two notes
Geocoding procedure	1–4 hours@$20/ hours		See step two notes
Reference street maps	Free	www.census.gov/geo/www/cob/bdy_files.html TIGER 1990, 1995 & 2000 maps available	See step two notes. County. zip code, incorporated places, etc. boundary files are available from the US Census Bureau

CASS = coding accuracy support system for mailing addresses; HIPAA = health insurance portability and accountability act; IRB = institutional review board; ESRI = Environmental Systems Research Institute, Inc; TIGER = topologically integrated geographic encoding and reference system; GDT = Geographic Data Technology company.

* Estimated costs for our project (geocoding ≈15,000 addresses).

† Approximately 3,500 problematic addresses.

ers or study participants. Tax assessors' books can be used to determine the parcel of land corresponding to street addresses. Locations can be transferred to United States Geological Survey topographical maps and then digitized.[24–26] Alternatively, a qualified land surveyor can visit each site to identify locations.[27] Although the latitude/longitude coordinate assignment rate can approach 100%, these methods are expensive and would not be practical in studies with large numbers of pre-existing addresses or that involve expansive geographic areas. For studies that cover small geographic regions, street reference maps available from a local agency can provide better match rates than TIGER street maps[32,33] and can be less expensive than street reference maps from commercial vendors.[34]

Other methods of improving match rates include using participants' nearest intersections[4] to geocode or using Internet directories[20] to update street addresses. We found intersections were twice as likely as addresses to geocode (step 4). However, the effectiveness of the intersection information varied according to the completeness of Wisconsin's TIGER street map, especially in rural locations. In addition, addresses from Internet directories may be inaccurate, since individuals may move within a city but retain the same telephone number.

Studies commonly report or imply matching rates without providing information about sensitivity scores or geocoding methods.[35–39] Because of this omission, critical evaluation of the study, as recommended by Drummond,[34] is limited. For example, in our study the original 76% match rate masked high regional variability, particularly for rural and urban areas.

Although the match rate is important, accuracy of the geocode is important as well. The potential limitations of geocoding using street reference maps must be considered. Street reference maps contain some degree of positional inaccuracy, which represents the distance between true and assigned coordinates.[40] An example of positional inaccuracy would be a geocode assigned to the wrong coordinates although the sensitivity score may be high. Because ArcView uses point-in-polygon geometry for census tract assignments, addresses (points) that are placed on street centerlines and that coincide with tract boundaries (polygons) can be incorrectly assigned. Rushton advises adding a small positive value to the street offset parameter in ArcView, especially when researchers are intending to use the geometrical location to find administrative areas (personal written communication from Gerald Rushton, University of Iowa, 24 December 2002). Another source of misclassification is the temporal nature of street reference maps. Inspecting, updating, and validating changes for a statewide street reference map cannot occur simultaneously.[41] The degree of positional accuracy of the TIGER street maps used in the geocoding process outlined in this paper is not known, although a high level of precision is shown in our study by a six-digit latitude/longitude coordinate assignment.

A variety of street reference maps are available for geocoding. Map vendors improve the TIGER street maps by using US Postal Service, emergency 911, and local agency data.[6] There are disadvantages of the enhanced maps. For example, the source documentation may not exist or may be uncertain. Commercial licensing may be restrictive on end-product use. Also, the maps are relatively expensive.[42] In contrast, TIGER street maps are free and source documentation describing the files is readily available. TIGER street maps have near universal compatibility with geocoding software,[11] although the Census Bureau cannot guarantee completeness.[6] Disadvantages of TIGER street maps include infrequent updates, and reliance on state and sometimes local government agencies (who often have limited budgets) to provide the source data.

Even though the positional accuracy of the geocoded addresses using different street reference maps has yet to be quantified, the importance of geocoding at the street address level is exemplified in studies by Krieger et al[43] and Lim et al.[44] Krieger et al demonstrated that zip code-level analysis using geocoded data yielded results for colon cancer incidence that were contrary to census tract or block group level analysis.[43] In a study of Pennsylvania cancer registry records, geocodes corresponding to county and municipal codes were compared with geocodes for mailing addresses.[44] These two methods resulted in substantial differences in breast cancer incidence rates by county.

In conclusion, geocoding can be useful for evaluating community-level factors,[45] assigning exposure levels of an environmental toxin,[46] and determining distance to health care resources.[47] However, researchers must have a thorough understanding of the various components of the geocoding process to geocode study data effectively and assess the quality of geocoded addresses. Exploring, evaluating, and describing different geocoding methods will assist in adoption of the best practices for selecting appropriate geocoding methods.

ACKNOWLEDGMENTS

We are indebted to all women who participated in our studies. We are grateful to Drs. Henry Anderson, Larry Hanrahan, Russell Kirby, Marty Kanarek, Colin Jefcoate, and William Sonzogni for advice and support on this project; Laura Stephenson of the Wisconsin Cancer Reporting System for assistance with data; Betty Granda, Christina Kantor, Elizabeth Mannering, Kathy Peck, Lisa Sieczkowski, Jerry Phipps, John Hampton, Nicole Angresano, Mina Kim, and Linda Haskins for data collection and study management; Ayak Reec, Jeffrey Pearson, Indiana Strombom, Stephanie Holmes, LeAnn Anderson, Kwang Kim, and Luxme Harihan for geocoding; and Mary Pankratz, Math Heinzel, Peter Nepokroeff, John Laedlein, Amy Sapp, Michael Kantor, and Gene Hafermann for technical support.

REFERENCES

1. Rushton G, Krishnamurthy R, Krishnamurti D, et al. The spatial relationship between infant mortality and birth defect rates in a U.S. city. *Stat Med*. 1996;15:1907–1919.
2. Krieger N, Chen JT, Waterman PD, et al. Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter? The Public Health Disparities Geocoding Project. *Am J Epidemiol*. 2002; 156:471–482.
3. Rushton G. Methods to evaluate geographic access to health services. *J Public Health Manage Pract*. 1999;5:93–100.
4. Vine MF, Degnan D, Hanchette C. Geographic information systems: their use in environmental epidemiologic research. *Environ Health Perspect*. 1997;105:598–605.
5. Daniel L, Slezak J. Street talk: The word on address matching. *Business Geographics*. 1995:1–11.
6. Johnson SD. Address matching with commercial spatial data: part 1. *Business Geographics*. 1998:24–32.
7. Johnson SD. Address matching with stand-alone geocoding engines: part 2. *Business Geographics*. 1998:30–36.
8. Krieger N, Waterman P, Lemieux K, et al. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. *Am J Public Health*. 2001;91:1114–1116.
9. Taylor D, Chavez G. Small area analysis on a large scale–The California experience in mapping teenage birth "hot spots" for resource allocation. *J Public Health Manag Pract*. 2002;8:33–45.
10. MacDorman MF, Gay GA. State initiatives in geocoding vital statistics data. *J Public Health Manag Pract*. 1999;5:91–93.
11. Lee CV, Irving JL. Sources of spatial data for community health planning. *J Public Health Manag Pract*. 1999;5:7–22.
12. Thrall SE. Geographic information system (GIS) hardware and software. *J Public Health Manag Pract*. 1999;5:82–90.
13. Newcomb PA, Storer BE, Longnecker MP, et al. Lactation and a reduced risk of premenopausal breast cancer. *N Engl J Med*. 1994;330: 81–87.
14. Newcomb PA, Egan KM, Titus-Ernstoff L, et al. Lactation in relation to postmenopausal breast cancer. *Am J Epidemiol*. 1999;150:174–182.
15. Marx RW. The TIGER system: automating the geographic structure of the United States Census. In: Peuquet DJ, Marble DF, eds. *Introductory Readings in Geographic Information Systems*. London: Taylor and Francis: 1990:120–141.
16. ESRI. Summary of controlling match weights and computing scores in ArcView geocoding: ESRI, 2002.
17. Eberhardt MS, Ingram DD, Markuc D. *Urban and Rural Health Chartbook*. Hyattsville, MD: National Center for Health Statistics; 2001.
18. Census Bureau US. Federal Register. Washington, DC: U. S. Government Printing Office; 2002:11665–11670.
19. McLafferty S, Cromley E. Your first mapping project on your own: from A to Z. *J Public Health Manag Pract.* 1999;5:76–82.
20. Boscoe FP, Kielb CL, Schymura MJ, et al. Assessing and improving census tract completeness. *J Registry Manage*. 2002;29:17–20.
21. Diez-Roux AV, Nieto FJ, Caulfield L, et al. Neighbourhood differences in diet: the Atherosclerosis Risk in Communities (ARIC) Study. *J Epidemiol Community Health*. 1999;53:55–63.
22. Chen FM, Breiman RF, Farley M, et al. Geocoding and linking data from population-based surveillance and the US Census to evaluate the impact of median household income on the epidemiology of invasive *Streptococcus pneumoniae* infections. *Am J Epidemiol*. 1998;148:1212–1218.
23. Hwang SA, Fitzgerald EF, Cayo M, et al. Assessing environmental exposure to PCBs among Mohawks at Akwesasne through the use of geostatistical methods. *Environ Res*. 1999;80:S189–S199.
24. Timander LM, McLafferty S. Breast cancer in West Islip, NY: A spatial clustering analysis with covariates. *Soc Sci Med*. 1998;46:1623–1635.
25. Paulu C, Aschengrau A, Ozonoff D. Exploring associations between residential location and breast cancer incidence in a case-control study. *Environ Health Perspect*. 2002;110:471–478.
26. Kohli S, Brage HN, Lofman O. Childhood leukaemia in areas with different radon levels: a spatial and temporal analysis using GIS. *J Epidemiol Community Health*. 2000;54:822–826.

27. Rajput AH, Uitti RJ, Stern W, et al. Geography, drinking water chemistry, pesticides and herbicides and the etiology of Parkinson's disease. *Can J Neurol Sci*. 1987;14:414–418.
28. Rushton G, Lolonis P. Exploratory spatial analysis of birth defect rates in an urban population. *Stat Med*. 1996;15:717–726.
29. Ross A, Davis S. Point pattern analysis of the spatial proximity of residences prior to diagnosis of persons with Hodgkin's disease. *Am J Epidemiol*. 1990;132:S53–S62.
30. Fulcomer MC, Bastardi MM, Raza H, et al. Assessing the accuracy of geocoding using address data from birth certificates: New Jersey, 1989 to 1996. GIS in Public Health Conference. San Diego, CA, 1998:547–560.
31. Davis S, Ross A, Voigt LF, Heuser L. Qualitative and quantitative assessment of geographic clustering of population samples selected using different methods of random digit dialing. *Am J Epidemiol*. 1990;132:S144–S155.
32. Rogers MY. Getting started with Geographic Information Systems (GIS): a local health department perspective. *J Public Health Manag Pract* 1999;5:22–33.
33. Lang L. *GIS for Health Organizations*. Redlands, CA: ESRI, Inc; 2000.
34. Drummond WJ. Address matching: GIS technology for mapping human activity patterns. *J Am Planning Assoc* 1995;61:240–251.
35. Xiang H, Nuckols JR, Stallones L. A geographic information assessment of birth weight and crop production patterns around mother's residence. *Environ Res* 2000;82:160–167.
36. Diez Roux AV, Merkin SS, Arnett D, et al. Neighborhood of residence and incidence of coronary heart disease. *N Engl J Med* 2001;345:99–106.
37. Sheehan TJ, Gershman ST, MacDougall LA, et al. Geographic assessment of breast cancer screening by towns, zip codes, and census tracts. *J Public Health Manag Pract* 2000;6:48–57.
38. Margai F, Henry N. A community-based assessment of learning disabilities using environmental and contextual risk factors. *Soc Sci Med* 2003;56:1073–1085.
39. Cohn P, Klotz J, Bove F, et al. Drinking water contamination and the incidence of leukemia and non-Hodgkin's lymphoma. *Environ Health Perspect* 1994;102:556–561.
40. Drummond J. Positional accuracy. In: Guptill SC, Morrison JL, eds. *Elements of Spatial Quality*. Oxford: Elsevier Science Ltd; 1995:31–58.
41. Guptill SC. Temporal information. In: Guptill SC, Morrison JL, eds. *Elements of Spatial Quality*. Oxford: Elsevier Science Ltd; 1995:153–166.
42. Richards TB, Croner CM, Rushton G, et al. Geographic information systems and public health: mapping the future. *Public Health Rep* 1999;114:359–360.
43. Krieger N, Waterman P, Chen JT, et al. Zip code caveat: bias due to spatiotemporal mismatches between zip codes and US census-defined geographic areas–the Public Health Disparities Geocoding Project. *Am J Public Health* 2002;92:1100–1102.
44. Lim ST, Spielberg LA, Dolen MA. Controlling for a proximate determinant of breast cancer—a creative use of geographic information systems (GIS) for analyzing data with sparse background information. National Conference on Health Statistics. Washington DC, 1999:1–14.
45. MacIntyre S, Ellaway A, Cummins S. Place effects on health: how can we conceptualise, operationalise and measure them? *Soc Sci Med* 2002; 55:125–139.
46. Brody JG, Ruthan R. Mapping out a search for environmental causes of breast cancer. *Public Health Reports* 1996;111:494–508.
47. Andersen R. The multiple and changing faces of access. *Med Care* 1998;36:252–253.